DR.G.Y.PATHRIKARCOLLEGEOFCOMPUTERSCIENCEAND INFORMATION TECHNOLOGY, N-6, CIDCO, ChhatrapatiSambhajinagar,Maharashtra,India

SynopsisReport

On

Design&AnalysisforEfficientRetrievalofText Summarization

Submittedby

Ms.DeepaliVaijinathSawane

ResearchScholar
Dr.G.Y.PathrikarCollegeofComputerScience&IT MGM UNIVERSITY

Guidedby

Dr.SanjayY.Azade

AssociateProfessor
Dr.G.Y.PathrikarCollegeofComputerScience&IT MGM UNIVERSITY

Ph.D(ComputerScience)SixthSemester, Academic Year 2023-2024

DR.G.Y.PATHRIKARCOLLEGEOFCOMPUTERSCIENCEAND INFORMATION TECHNOLOGY, N-6, CIDCO, ChhatrapatiSambhajinagar,Maharashtra,India

CERTIFICATE

This is to certify that, **Ms. DEEPALI VAIJINATH SAWANE** has successfully completed the Synopsis Report on **"Design & Analysis for Efficient Retrieval of Text Summarization"** for partial fulfillment ofthe Program Ph.D (Computer Science) Sixth Semester, of MGM University, Dr. G. Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar during the Academic Year 2023- 2024.

Dr.SanjayY.Azade
ProjectGuide

Dr.SanjayY.Azade
HeadoftheDepartment

Examiner

TABLEOFCONTENTS

LISTOF FIGURES iv

LISTOF TABLES iv

1 INTRODUCTION 6

TYPESOFDATA SUMMARIZATION 7

TEXTSUMMARIZATION 8

GRAPHSUMMARIZATION 8

IMAGESUMMARIZATION 8

VIDEOSUMMARIZATION 8

AUDIOSUMMARIZATION 9

SUMMARIZATIONTECHNIQUES 9

BASEDON APPROACHES 9

BASEDONTYPEOF DETAILS 9

BASEDON TYPEOF CONTENT 10

BASEDON LIMITATION 10

BASEDONNUMBEROFINPUT DOCUMENT 10

BASEDON LANGUAGE 10

TYPESOF SUMMARIZATION 10

DOCUMENTBASEDSUMMARIZATION 11

GENERICORDOMAINSPECIFICSUMMARIZATION.....11

QUERY-BASED SUMMARIZATION.....11

EXTRACTIVETEXTSUMMARIZATION 12

ABSTRACTIVETEXTSUMMARIZATION.....12

HOWSUMMARIZATIONPLAYSANIMPORTANTROLE 12

APPLICATIONOFTEXT SUMMARIZATION.....13

NewsArticle Summary:.....14

EmailSummary:.....14

Shortmessageofnews on mobile:.....14

Informationsummaryfor businessman:.....14

Governmentofficials:.....14

Researchers: 15

MedicalRecords: 15

MAJORCHALLENGESINTEXTSUMMARIZATION.....15

RESEARCHOBJECTIVES.....15

2 LITERATUREREVIEW.....16

3 PROPOSEDMETHODOLOGY 21

OVERVIEWOFTHE NARRATIVES.....21

SOURCEMANUSCRIPT(NARRATIVES).....22

PRE-PROCESSINGTOOLS.....23

SENTENCETOKENIZATION.....23

WORDTOKENIZATION.....24

POS Tagging.....24

RuleBasedTaggers:.....25

StochasticTaggers.....25

Hybrid Taggers:.....25

NamedEntityRecognition.....25

FORMATIONOFRULES.....26

FEATUREEXTRACTION.....26

Text-RankAlgorithm.....27

RELATIONBETWEEN EXTRACTEDTEXT.....27

4 CONCLUSION:.....28

REFERENCES.....29

PUBLICATIONS:.....32

1 INTRODUCTION

The world today is rapidly growing as a vast repository of information technology, with most of it readily accessible on the World Wide Web. The Internet contains billions of records, and this number is growing exponentially (Ayush Agrawal et. al., 2014). Text summarization is atechniquethat reduces thequantity oftext in thesource document to a condensed version without altering itsmeaning. It effectively retains the essential content of the original articles.With the rapid increase in the internet usage, searching for information becomes a tedious and time-consuming task due to the overabundance of the results available (Al-Hashemi, 2010). Text summarization is crucial for various applications,includingsearchengines,businessanalysis,marketreviews,electronicmail, tracking the history of patient's medical records, scientific blogs, spontaneous dialogues, voicemail,news andbroadcastvideo,andmeetingrecordings (Deepali K. Gaikwad et.al., 2016).Natural Language Processing (NLP) is an important domain within computer science and artificial intelligence that focuses on how machines understand human language. Text summarization, a growing subfield of NLP, aims to condense documents into their shorter versions that retain their essential information. This process addresses the user's need for brief, yet comprehensive content (P.S.Kamble et al., 2017). The proliferation of online data, fueled by rapid digitization and internet expansion, provides abundant textual material for this purpose. The development and accessibility of online data are central to the research in text summarization within NLP (Dipanjan Das et al., 2007). The sheer volume of available information leads to a phenomenon termed "Infobesity" (V.Gupta et al., 2010), which describes the overwhelming amount of data accessible online. This makes selecting and processing relevant information challenging. Text summarization effectively mitigates this challenge by presenting concise summaries of documents, allowing users to efficiently access specific information without the needto sift through extensive texts (C. P. Sumathi et. al., 2012). These concerns have spurred interest in the expansion of automatic summarization systems. These systems are specifically designed to handle various types of content, such as editorial collections, groupsofnewsarticles,andbroadcastnewscompilations,creatingsynopsesthat

highlight the most essential information (Eduard Hovy et. al., 1999). Text summarization enables users to quickly access information, making it easier to compile important details from multiple sources into a single document. Techniques like keyword extraction are employed to identify significant keywords from documents. Keyword and feature extraction are fundamental challenges in text data mining and document processing, with applications spanning website comprehension, mobile messaging summarization, and email categorization (Ramond J. Mooney et.al., 2005).News groups can utilize multi-document summarization to efficiently gather and collate information (Nimisha Dheer et.al., 2013). We are receiving only relevant text instead of reading tedious and baffling text. This is an application of Natural Language Processing and it is one of the essential tools for understanding textual information.

“A summary is a manuscript created from multiple source texts, conveying a significant portion of the material in the source texts, and is no longer than half or one-third of the source texts' length” (Nenkova, Ani, 2011). This simple definition highlights three important features that characterize research on automatic summarization (Kamble, Mandage et. al., 2017).

- Summariescanbeproducedfromasingledocumentormultiple documents.
- Thesummariesshouldpreserveallimportantinformation.
- Thesummaries shouldbe concise.

Automatic text summarization systems select important sentences from a document and group them together in a matter of minutes, enabling users to quickly understand the content within the larger document. The purpose of automatic text summarization is to condense a large amount of data into short, essential content (Ayush Agrawal et.al., 2014). To identify key sentences for summarization, features such as title, location, similarity, proximity, headline, cue method, relation, nouns, verbs, adjectives, and sentence length are all utilized.

TYPESOFDATA SUMMARIZATION

Datasummarizationcanbeconductedusingvariousapproachessuchastext

summarization, graph summarization, image summarization, video summarization and audio summarization.

TEXTSUMMARIZATION

Data summarization can utilize text for summary generation. The text summarization process involves reducing the size of the original document to a condensed, smaller version, producing a brief and fluent summary without altering its meaning. Text summarizationisclassifiedintoseveraltypes:document-basedsummarization,genericor domain-specific summarization, query-based summarization, extractive text summarization, and abstractive text summarization (AyushAgrawal et.al., 2014)..

GRAPHSUMMARIZATION

Graphsareverypopular insocialnetworks,theweb,communicationnetworks,andsemi-structured data. Graph summarization explains the relationships between objects and edges instead of just listing them. It assigns a set of characterizing attributes to each element, such as age, location, relation, and function. The main purpose of the graph summarizationprocessistomaketheinformationingraphicsaccessibletosightlessusers through keyboard instructions. It uses templates to provide a brief textual summary of whatthegraphiclookslike,butthesesummaries donotconveythe high-levelknowledge provided by the graphics (Nimisha Dheer et. al., 2013).

IMAGESUMMARIZATION

Image summarization involves creating a summary of a text document using images. Image processing is a method used to convert an image into digital form and perform operations on that digitalimage to enhance it and extract useful data or points (Prabhu et. al., 2018).

VIDEOSUMMARIZATION

This method represents a conceptual analysis of creative videotapes and is also used in video browsing and retrieval systems. It can emphasize the actual video, which is a combinationofuser-definedsegmentsoracollectionofkeyframes.Thevast

development of digital video capture and control of editing equipment has led to an increase in video recording data, necessitating the development of effective video summarization techniques (Kamal Sarkar et. al., 2011).

AUDIOSUMMARIZATION

Audio summarization involves voice recordings, musical soundtracks, sound effects, audio descriptions, and real-time two-way communication. It is a powerful tool for learning materials and environments. Audio summarization is the task of reducing thesize of a source audio clip to short and important points without changing the essence of the actual audio clips. It is also very helpful in engaging students with their subjects inthe classroom (Duxans et. al., 2009).

SUMMARIZATIONTECHNIQUES

In addition to abstractive and extractive text summarization, there are various other types of summarization. These techniques can be useful in many different applications and can be classified according to approaches, detail types, content types, limitations, the number of input documents, and languages, as follows:

BASED ON APPROACHES

Extractive text summarization and abstractive text summarization are two strategies for summarization. Extractive text summarization involves selecting sentences directly from the source text and incorporating them into the summary. Abstractive Summarization, on the other hand, involves creating new sentences to provide a synopsis (Nimisha Dheer et. al., 2013).

BASEDON TYPEOFDETAILS

Thetypeofdetailssummaryisusedtoprovideaconciseoverviewofalengthydocument withoutchangingthecoreideasoftheoriginaldocument.Typesofdetailssummariescan be either informative or indicative. This kind of summarization enhances a user's interest in reading the original document (Rafeeq Al-Hashemi et. al., 2010).

BASED ON TYPE OFCONTENT

In the type of content summary, all information retains the level of importance of the original document and is not user-specific. The generic summarization system is user-friendly and depends on the original document but presents it in its own way (Rafeeq Al-Hashemi et. al., 2010).

BASED ON LIMITATION

The summary type based on limitations accepts specific types of input text such as newspapers, articles, narratives, and manuals, among others. It does not accommodate general types of documents. This approach is limited to certain specific types of input, generally topic- or domain-specific (Rafeeq Al-Hashemi et. al., 2010).

BASEDONNUMBEROFINPUT DOCUMENT

In the type of summary based on the amount of input documents, there can be either single-document summarization or multi-document summarization. Single-document summarization involves creating a summary from one document only. Multi-document summarization, on the other hand, involves summarizing multiple documents. Creating a summary from a single document is generally easier than producing a summary from multiple documents, as multi-document summarization presents additional challenges in integrating and synthesizing information across various sources (Siva kumar et. al.,2011).

BASED ON LANGUAGE

The type of summary based on languages can be classified into monolingual and multilingual systems. A monolingual system processes documents in a single language.In contrast, a multilingual system accepts and processes documents in multiple languages (Bhavna Lanjewar, 2015).

TYPESOF SUMMARIZATION

Summarization can be classified into various types depending on the method used to createasummaryofthedocument'stext.Textsummarizationmethodscanbecategorized into several types, including document-based summarization, generic or domain-specific summarization, query-based summarization, extractive text summarization, and abstractive text summarization (Dipanjan et. al., 2007).

DOCUMENTBASEDSUMMARIZATION

Document-based summarization is of two types: single-document summarization and multi-document summarization. Summarization can be classified based on the type of inputithandles.Single-documentsummarization createsasummarythatencapsulatesthe relevant data from a single input document. This helps users decide whether thedocument is pertinent to their topic of interest. Multi-document summarization, as the name suggests, creates a summary by synthesizing multiple input documents. This typeof summary is useful for identifying repetitions in documents and compiling summaries of related documents in a corpus, covering the major details of the situations described in the documents. It often deals with similar topics, such as several news articles. These are processed individually, and the output summaries may contain a considerable amount of repeated information due to overlap in the source materials (Richa Sharma, 2017).

GENERICORDOMAINSPECIFICSUMMARIZATION

Summarization can be classified based on purpose. In generic or domain-specific summarization, the aim is to summarize all documents or texts related to a specific domain. Generic summarization does not focus on any particular group; instead, it addresses a broad audience. In this process, the essential task is to identify terms and phrases specific to the domain (Kamal Sarkar et.al.,2011).

QUERY-BASEDSUMMARIZATION

Query-basedsummarizationincludesonlythedatarequiredbytheuser.Aquery-based

document summarizer operates based on the similarity of sentences and word frequency. The queries, which are natural language questions, pertain to a specific topic. Query-based summarization is similar to a question-answering task in that it extracts words relevant to the user's query in response to their input (A.P. Siva Kumar et. al., 2011).

EXTRACTIVETEXTSUMMARIZATION

Summarization can be classified based on the type of output it produces. An extractive summary selects important sentences, which are identified directly from the document. Extractive text summarization gathers crucial information from the original data and groups it together without altering its meaning. The methods used in extractive text summarization include the term frequency-inverse document frequency (TF-IDF)method, cluster-based method, graph-theoretic approach, machine learning approach, Latent Semantic Analysis (LSA) method, text summarization with neural networks, and automatic text summarization based on fuzzy logic (Bhavana Lanjewar et.al., 2015).

ABSTRACTIVETEXTSUMMARIZATION

Abstractive text summarization is performed by understanding the original text with the help of linguistic methods. The main purpose of abstractive summarization is to developa relevant summary that accurately conveys information, often requiring advanced language generation techniques. In abstractive summarization, selected sentences fromthe document are systematically concatenated and reduced to omit unnecessary text. This method requires a deeper analysis of the text document. Abstractive summarization is classified into structured and semantic approaches. The structured approach is based on extracting essential material from the source using models, extraction rules, and other structures such as trees, ontologies, lead phrases, and body phrases (Roshna Chettri,et.al., 2017).

HOWSUMMARIZATIONPLAYSANIMPORTANTROLE

Summarization has become increasingly popular among researchers, readers, and innovativeprojectideas.Ithelpsresearchstudentslearntoidentifyessentialideasand

consolidate important details that support them. It enables students to focus on keyword phrases of an assigned text that are worth noting and remembering (YashashviSharma et. al., 2017).

On the internet, there are various examples of summarization tools such as news and article summarizers like Microsoft News2, Google1, FreqDist, Sum Basic, MEAD, and AutoSummarize. SWESUM is a popular biomedical summarization tool. Text Compactor, Simplify, Tools4Noobs, Free Summarizer, Wiki Summarizer, andSummarize Tool are among the online summarization tools available.

The future task in automatic text summarization involves using extractive methods to extract significant and efficient features and key phrases from specified textual documents. The proposed technique aims to produce summaries that incorporate various external sources and process the input using preprocessing tools such as sentence and word splitting, tokenization, Part of Speech (PoS) tagging, anaphora resolution, text ranking algorithms, and Named Entity Recognition. These tools are intended to generate high-quality summaries based on extractive text summarization techniques.

APPLICATIONOFTEXTSUMMARIZATION

Summarizationcanbeneededforvarious purposes,such as:

Figure1:ApplicationofTextSummarization

NewsArticle Summary:

A summary offers users or readers an opportunity to understand the important ideas of a topic without the need to read the original, lengthy document. News articles may be long and can take too much time to get to the actual text and points (Deepali k. Gaikwad et.al.,2016).

EmailSummary:

Email is short for Electronic Mail. It's a medium of communication through mail. It involves the process of exchanging messages between people using electronic devices.Anemailsummary ishelpfulforreaderswhoarereceivingthemainideasofalongemail as a summarized version (Deshpande, et. al., 2018).

Shortmessageofnewson mobile:

Weekly news often takes the form of an introduction to relevant articles. Summarization helps enrich the stream of news, making it a suitable format for mobile devices (Bharti, Babu, & Pradhan, 2017).

Informationsummaryforbusinessman:

A summary is very useful for businessmen, providing valuable key points from original andlengthydocumentsinlesstime. Italsohelps businessmen decideon meetingagendas and business deals (Prabhu et al., 2018).

Government officials:

A summary is very useful for businessmen as it provides key points from original and lengthy documents quickly. It also assists in deciding the agendas for meetings and business strategies (D.K. Gaikwad et. al., 2016).

Researchers:

The abstract of a research paper is a human-made summary. When researchers are monitoring trends and innovations in a particular sector, it can be very difficult to read every abstract. Therefore, a system that can collect and compress these abstracts wouldbeveryhelpfulforresearchers.Inthefieldofresearch,reviewingpatentsisalsoatedious task; thus, a summarizer that highlights salient claims can be particularly beneficial (A. Deshpande et. al., 2018).

Medical Records:

In the medical field, when a patient consults a doctor about their health issues, the doctor quickly gathers the patient's history during their first meeting (Gaikwad et. al., 2016).

MAJOR CHALLENGESINTEXTSUMMARIZATION

- Themajorproblemisselectingthedatathataresignificant.
- Extractivetechniquesmaybeineffectiveinexpressingtherelationshipsbetween methodologies in a text.
- Themost importantsentences inthesummaryshouldbewithoutredundancy.
- Evaluating a summary isa complicated process because there is no absolute summary for a given document or set of documents.
- Theuseofsummary evaluationiswidespreadand misusedacrossdifferentmatrices.
- Timeisveryimportantforthereader;therefore,textsummarizationhelpstheusertoread the maximum amount of text document in the minimum time without changing the essence of that text document.
- This is a small effort to compress many different documents into a single document with important information.

RESEARCHOBJECTIVES

Togenerateextractivetextsummarizationfollowingobjectivesareset:

- Toframe rules ontheassociative-based systemfortext identification.
- Toapplyextractivetechniquesforfeatureextraction.
- TodevelopAnaphoraResolution& NamedEntityRelationwithrespecttothenarratives.
- To capture association of sentences using anaphora resolution to accomplish consistency in the summary.
- Toimprovetheperformanceofsummary.

2 LITERATUREREVIEW

Table1:LiteratureReviewon Text Summarization

Author& Year	Techniques/Methods	Description	Lacuna
M.Fachrurrozi, Novi Yusliani, and Rizky UtamiYoanita, (2013)	Term Frequency Inverse Document FrequencyMethod	Sentence frequency is defined as thenumbers of highest scoring sentences that arepicked to be part of the summary.	There is necessary ofthe system can be improved by integrating the system with another technique to get the enhanced summary. Besides being able to generate extraction the text summarization can also produces abstractions, the explanationofthe originaltext.
Mehdi Bazghandi, GhamarnazTadayon Tabriziand Majid VafaeiJahan (2012), Anjali R. Deshpande, LoboL. M.R.J.(2013)	Cluster-basedmethod	It is intuitive to think thatsummariesshould address different “themes”appearingin the documents. If the documentcollection forwhichsummaryis	Limitation is thatall themethodsof summarization are done with the 50% only.

		being produced is of totally different topics, document clustering becomes to generateameaningful summary	
Rada Mihalcea, NirajKumar,Kannan Srinathanand VasudevaVarma, (2013), Sarda A.T. andKulkarni A.R. (2015)	GraphTheoretic Approach	Graph theoretic representation of passages provides a method of identification of themes. After the common pre-processing steps, namely, stemmingand stop word removal; sentences in the documents are representedasnodesin anundirected graph.	There is limited linguistic support to include only stop words, stemmer and punctuation marks to achieve the goal.
Patil PallaviD,Mane PM(2014), Hanane Froud,Abdel monaimeLachkarand SaidAlaoui Ouatik (2013), Mr. S.A.BabarandProf. S.A.Thorat(2014	LSAMethod	It gets this name LSA because SVD applied to document word matrices, group documents that are semanticallyrelatedto each other, evenwhen they do not share commonwords.	There is a necessity of improving the techniques for obtaining a better quality of summary by incorporating the latent semantic analysis into the sentence feature enhanced fuzzy logic system to capture the semantic relations betweenconceptsinthe

			text document.
Khosrow Kaikhan(2004)	Text summarization with Neural Networks	This method involves training the neural networks to learnthe typesofsentencesthat should be included in the summary. It uses three - layered Feed Forward neural network	There is highprobability that the article did not have subtitle or section heading.
Ms. Pallavi D.Patil, Prof. N.J.Kulkarni (2014),PatilPallavi D.,ManePM(2014) , Rucha S. Dixit,Prof. Dr. S. Apte (2012)	AutomaticTSbased on fuzzy logic	Thismethodconsiders each characteristic ofa text such as similarity to title, sentence length and similarity to key word etc.astheinputofthe fuzzysystem.	There is a necessity to use fuzzy logic as a summarization sub-taskfor developingan excellent of summary
Ibrahim Imam, Nihal Nounou, Alaa Hamouda,HebatAllah Abdul Khalek (2013),A.P.Siva kumar, Dr. P. PremchandandDr.A. Govardhan(2011)	QueryBased Extractive Text Summarization	In query - based text summarization system, the sentences in a given document are scored based on the frequency counts of terms. It uses Vector Space Model.	There is a need to improvethesystemby addingsentence simplification and Techniques for simplifying the sentences which are very large & complex.
SergeyBrin(1998)	Knowledge Based Method	Knowledge based is the first category of methods which speciallyusedin domain-specifictasks.	Knowledge-based methods are not easily accessible to another domain.

		In the knowledge based method the text are similarand closed set of relations needs to be identified. Inthismethodrelations atedomaindependent.	
Aron Culotta, Andrew McCallum, JonathanBetz(2006), Nanda Kambhatla (2004)	SupervisedMethod	Supervisedmethodsare automatic extract therelation by using Machine-learning Techniques. This method is domain Specific. This method can easily accessedon anotherdomain.	The limitation of this methodis the developmentof suitable tagged can take lot of time and effort.
ZornitsaKozarevaand EduardHovy(2010), Michele Banko and Oren Etzioni(2008), Eduard Hovy and Zornitsa Kozareva, Ellen Riloff(2009)	Weakly-Supervised Method	Weakly-Supervised Methodisalsousedin bootstrapping tomake construction ofdata easily. This technique is also called "Weakly-SupervisedInformation Extraction."	The serious problem of this method is Propagation and error to expand named Entity Recognition extracting in Incomplete name.
Natalia Konstantinova(2014), Michele Banko and Oren Etzioni(2008)	Self-Supervised Method	Self-Supervised Method go further task of information Extraction unsupervised.This methodusessetof generic pattern to automatically	This method repeated all process iteratively.Themajor problem of this method is that all relation cannot be extracted.

		instantiate relation specific extraction rules.	
AbdelkrimeAries,et al (2015)	Stemmer,Stop-Words eliminator, Naive Bayes	In this technique sentence extraction methodcanbeapplied to any language without any changes, affording the pre-process step of the input language therefore they find a new technique to train thesystemotherthan using corpus.	Limitation in extractingmethods.
Eugene Agichtein, LuisGravano(1999)	BaselineMethod	Thismethodisbased onfrequency of co-occurrence of the organization andthe location.It reports thelocation that co-occursinthesameline witheachorganization most often as the headquarters for organization.	Thismethodisable to exploit features.
Sergeybrin[1999]	Dual Iterative Pattern Relation Extraction (DIPRE) Method	This method used for extracting relation which makes use of pattern relation duality.Thismethod	Limitation of this method is that boththe task i.e. finding occurrencesofbooks andfindingpattern

		has two importanttask i.e. Firstly, findingoccurrencesof books and secondly finding pattern matches.	matches aretakeplace only large database of web document.
Eugene Agichtein& Luis Gravano(1999)	SnowballMethod	Snowball presents a novel methodology to Generate patterns and extract tuples from source text documents. Also, Snowball introduces a approach for evaluating the quality ofthe patterns and the tuples that are producedineach extractionmethod.	Thereisproblemwith snowballmethod,we do not trust,that tend to generate the wrong tuples because pattern is notselective, they have low weight.

3 PROPOSEDMETHODOLOGY

In this research work, we have analyzed some narratives from the World Wide Web and Panchtantra storybooks to generate a summary using automatic text summarization techniques.

OVERVIEWOFTHENARRATIVES

The repository of information has been significantly enriched by narratives. The relationshipbetweenfictionalorgenuineindividualsandoccurrencesstagedfor

entertainment constitutes the narrative. They enable us to convey situations in an understandable, emotionally empathetic manner. Additionally, narratives are important for intellectual and psychological development. According to psychologist G. Klein, novels are a vital educational technique because they reveal incidental connections that individuals might have never anticipated beforehand but that are helpful in explaining challenges. Reading summaries can save a ton of precious time compared to reviewing lengthy documents. When it comes to internet-based repositories that require vast volumes of statistics and records, a straightforward search function ought to produce a concise synopsis of the material in order to save the consumer's precious time. To implement the summarization of content approach, a large number of documents (short narratives) must first be entered. Phrases and sentences must subsequently be converted into tokens, POS taggers must be applied, and the recognition of named entities must be used to extract the key details and elements from the material or brief narrative, such as nouns, pronouns, verbs, adjectives, titles, actors, relations, places, morals, etc. The following lists the specifics of every system's phase:

SOURCEMANUSCRIPT(NARRATIVES)

We use a collection of literature (Ethical Narratives) as source material for the content summarization procedure. We selected 50 brief ethical narratives and child-friendly narratives from the World Wide Web. Each narrative has a minimum length of 19 sentences and a maximum length of 60 lines and sentences.

PRE-PROCESSINGTOOLS

One of the most important aspects of natural language processing, crucial to the content paraphrasing approach, is the pre-processing procedure. At this stage, the content of the manuscripts, including phrases, sentences, characters, spoken language, and features, is easily identifiable. Preliminary processing is the first stage in continuing processing after the evaluation and tagging of components such as information retrieval and morphological (Kamath, S. et. al., 2017). This set of exercises involves removing distinctive characters from the original material, such as exclamation points, punctuation

marks, and date formats, as well as reducing vocabulary like prepositions, nouns, documents, and pronouns. The preprocessing mainly involves phrase tokenization, POS labeling of the incoming written content, and removal of special characters (Devisree V. et.al. 2015). Numerous distinct characters, such as question marks, exclamation marks, and punctuation marks (", :, ;, ..., -), are used in narratives. These unique symbols are crucial for the reader's comprehension of each plot element. After that, the content is tokenized into words and POS tagged. (Schank, R. C. et. al., 2013).

SENTENCETOKENIZATION:

Theprimaryobjectiveoftheinitialstepsineve ryNLPworkflowistokenization.

Tokenization is simply the process of dividing the original text into tokens, which are discrete words or sentence fragments. Among the fields of natural language processing, also known as NLP, and Machine Learning (ML), tokenization, when possible, is the act of breaking down a sequence of texts into smaller components called tokens. These representations can range in length from words to characters. (Ayush Agrawal et.al., 2014). The character series those are significant for additional processing within any given group. The original document tokens are divided via this procedure.

WORD TOKENIZATION

Enormousstatementsarebrokenupintoasmallnumberofimportantparts oftheoriginal content using the phrase tokenization process. The tokens and sentences are separated by this procedure. If the phrases are split into sentences, then it's called Sentence Tokenization. The word tokenization procedure is a succession of characters in any meticulous compilation of words for further processing. The words are separated from some particular sentences and phrases.

POS Tagging

POS tagging is utilized in the pre-processing of the source material. A linguistic class of words is called a component of language (N. Jahangiret. et. al. 2011). This is animportant phase in language analysis, including the extraction of data, machine translation, along with data retrieval. For NLP programs, POS tagging is a desired preprocessing element. There was a pair of phases to the POS tagger separation. It begins by searching for keywords in the corpora; if the word is found, it is assigned a tag. If not, itderivesthewordingrammar,wheretherareeightpartsoflanguage (ByronGeorganto poulos etet.al. 2000). Similar to a noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection, a POS (Part of Speech) tagging is referred to as a token. In thisstudy,adjectives,nouns,andverbswereextracted fromtheoriginalmanuscriptusing POS tagging. The process of allocating words in a script that adheres to strict parts of speech requirements is known as Part-of-Speech (POS) tagging. It is a simple yet highly effective tool utilized in any natural language processing application. The reorganization ofthewordsasnouns,verbs,adjectives,etc.,providesacrucialreasonforPOStagging.

Rule-based tagging, stochastic tagging, and hybrid tagging are the three categories into which the various POS tagging approaches fall. POS labeling is necessary (G.F. Greebacker et.al., 2011). Additionally, the tagger is split into two groups: supervised and unsupervised. While unsupervised taggers automatically assign tags to words, supervised taggers rely on pre-tagged corpora. Rule-based tagging, stochastic tagging, and hybrid tagging are three categories of POS tagging approaches. (Sayyed S. N. et. al., 2018).

3.4.1 RuleBased Taggers:

A rule-based tagger is a type of Part-of-Speech (POS) tagger that assigns tags to words based on predefined linguistic rules. These rules are typically crafted by linguists or language experts and are based on grammatical patterns, word morphology, context, and syntactic structures.

3.4.2 StochasticTaggers:

Stochastic taggers are Part-of-Speech (POS) tagging systems that employ statistical models to assign tags to words based on probabilities derived from training data. Instead of relying solely on predefined rules or linguistic knowledge, stochastic taggers learn from annotated corpora to estimate the likelihood of a word being associated with a particular tag given its context.

HybridTaggers:

Hybrid taggers are Part-of-Speech (POS) tagging systems that combine elements of both rule-based and stochastic approaches to achieve improved tagging accuracy and robustness. These taggers leverage the strengths of each method while mitigating their respective weaknesses.

NamedEntityRecognition

When applying Named Entity Recognition (NER), issues arise regarding common nouns and proper nouns. Nouns are primarily used to recognize individuals, locations, objects, etc. However, there are challenges in distinguishing between common nouns and proper nouns.Forexample,"Friends"(NNS),"Sohan"(NNP),"village"(NN),and"bear"(NN)

may all refer to entities of different types, but they are all classified as nouns. It isdifficult to accurately classify these words; the system may fail to identify the proper noun type for a word. For instance, "Krishan Devraya" is a person's name and should be assigned the entity type "person," but in NER, it may be classified as an organization. While updating the corpus can resolve some of these issues, a significant challenge remains: common nouns in a story may erroneously be treated as proper noun.

FORMATIONOFRULES

The guidelines for an attribute of a source text are as follows: the title, the performer,their relationship, the activity occurring at a specific location and its explanation in sources, the performers' action, and above all, the ethical narratives. We have decided on a few final characteristics for a summary that are important in producing high-quality summaries and abstracting inspirational and ethical narratives. These are the framing of the parameters in rule-based.

FEATUREEXTRACTION

The feature extraction approach relies on POS tagging and uses extractive strategies to generate summary information by extracting features based on the Hidden Markov Model. The most important keywords from the original manuscript were recognized by the extraction algorithm, which then concatenated them into the abstract. The aforementioned five dimensions are used to illustrate the five fields of feature extraction. (Hamzah Noori Fejer et.al., 2014).

Thetypeofstructuredsourcemanuscript extracted:

- Thekindofunstructureddataorsource manuscript.
- Thecategoryofcontributionresourcemanuscriptcontentaccessible.
- Thestrategyisusedfortext,keywordandsentencesextractiontechnique.

- Theproductivityoffeatureextractionmethodology.

Thefivefieldofinformationextractionispresentedalong thesefivedimensions:

[A] Thetypeofstructured extracted.

[B] Thetypeunstructuredsources.

[C] Thetypeof Inputresourceavailable.

[D] Themethod used for extraction.

Text-RankAlgorithm

The Text-Rank system is an approach to machine learning that has resulted in a quadratic increasein theoutcomes oftext summarization methodology. Thetext-rank methodology is implemented to find the linked paragraphs and phrases in a particular set of provided content. These phrases are merged to provide a brief overview of the supplied content. Because it is a structure-based technique, the relevance of a vertex is determined by considering everything offered by a graph. It uses an unsupervised graph-based scoring algorithm for synopsis creation and performs the processes below to generate anoverview using source content.

RELATIONBETWEENEXTRACTEDTEXT

Regarding the text extraction from the narratives, we must determine which kind of feature (title, performer, position, relationship, and event) the extracted text corresponds to. We must compare the text with the feature in order to determine the text extraction feature.

4 EXPERIMENTALRESULTS

The created summary and abstract are based on the field of linguistics to provide the text summarizationprocedure.Bycontrastingthegeneratedsummarywiththebriefsummary, which is often produced by the linguistic field, the perspective of the system-derived summary is illustrated (kamath S. et. al. 2017). Typically, there are two categories for summary measurements: intrinsic and extrinsic. Average and intrinsically generated summariescanbecontrasted.Thesuperiorityofcreatingasummarybasedonhowit

complements the original abstract is verified in the extrinsic summary. We talk about the intrinsic techniques in terms of compression (Sayyed S. N.et. al., 2018).

5 CONCLUSION:

The vast growth in the rate of information due to the internet has called for the need for efficient text summarization systems. Text summarization plays a crucial role in dealing with the overwhelming volume of information available today. By condensing lengthytexts into concise summaries, text summarization tools help users save time and quickly grasp the main points of a document or article. With advancements in natural language processing and artificial intelligence, text summarization techniques continue to evolve, offering more accurate and efficient summarization capabilities. From extractive to abstractive methods,therearevarious approaches totextsummarization, eachwithits own strengths and limitations. This research work has provided a very easy overview of information extraction from any text document in various steps. Nowadays, many ready-made software options are available in the market such as text compactor, SMMRY, SummarizeBot, Resoomer and Text summarization etc. for the information extraction process, such as company names, employee details, locations, areas, relationships, characters, and activities, etc. The information extraction process is a very important technique for readers to get the important points of particular documents within seconds. Overall, text summarization has become an indispensable tool for researchers, students, professionals, and anyone else who needs to sift through large amounts of text efficiently. As technology continues to improve, we can expect text summarization to become even more sophisticated and effective in the future.

REFERENCES:

"E.Riloff,"InformationExtractionasaStepping StonetowardStoryUnderstanding.", In Understanding Language Understanding: Computational Models ofReading, edited by AshwinRam and Kenneth Moorman, The MIT Press.

Suneetha Manne, Zaheer Parvez Shaik Mohd. , Dr. S. Sameen Fatima, "ExtractionBased Automatic Text Summarization System with HMM Tagger", Proceedings of the International Conference on Information Systems Design and Intelligent Applications, 2012, Vol. 132, P.P 421-428.

" D. Sasirekha, E. Chandra ," Text Extraction from PDF document", *2013,* **.**Amrita International Conference of Women in Computing (AICWIC"13) Proceedings published by International Journal of Computer Applications® (IJCA).

"Nimisha Dheer, Chetan Kumar **,**2014,**"** Automatic Text Summarization: A Detailed Study"**,** International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

"Dipanjan Das Andre F.T. Martins, "A Survey on Automatic Text Summarization" November 21, 2007.

"Viswanath Meghana, "Thesis: Ontology-Based Automatic Text Summarization", M. Sc Thesis, Vishweshwaraiah Institute of Technology, India, 2009.

" Mehdi Bazghandi, Ghamarnaz Tadayon Tabrizi and Majid Vafaei Jahan, "Extractive Summarization of Farsi Documents Based On PSO Clustering", InternationalJournal of Computer Science Issues, 2012, Vol. 9, Issue 4, No 3.

Deshpande Anjali R., Lobo L. M. R. J., "Text Summarization using Clustering Technique", InternationalJournalofEngineering TrendsandTechnology(IJETT),2013, Vol. 4Issue8.

SivakumarA. P., Premchand P. and Govardhan A., "Query-Based SummarizerBased on Similarity of Sentences and Word Frequency", International Journal of Data Mining & Knowledge Management Process (IJDKP),2011, Vol.1, No.3.

Kumar Niraj, Srinathan Kannan and Varma Vasudeva, “A Knowledge Induced Graph-Theoretical Model for Extract and Abstract Single Document Summarization”, Computational Linguistics and Intelligent Text Processing - 14th International Conference,2013.

Bazrfkan Mehrnoosh And Radmanesh Muosa, “Using Machine Learning Methods To Summarize Persian Texts”, Indian J.Sci.Res. 7 Issue1: 1325-1333, 2014.

Josef Steinberger, Karel Je˘zek, “Evaluation Measures For Text Summarization”, Computing and Informatics, Vol. 28, pp 1001–1026, 2009.

Sarda A.T. and Kulkarni A.R., “Text Summarization using Neural Networks and RhetoricalStructure Theory”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, 2015.

Subramaniam Manjula, Dalal Vipul, “ Test Model for Rich Semantic Graph Representationfor Hindi Text using Abstractive Method.” , IRJET, 2015.

Oya Tatsuro, Mehdad Yashar, Carenini Giuseppe, and Ng Raymond, “A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships”, Proceedings of the 8th International Natural Language Generation Conference, pages 45–53,Philadelphia, Pennsylvania, 2014.

Patil Pallavi D., Kulkarni N.J., “Text Summarization Using Fuzzy Logic”, International Journal of Innovative Research in AdvancedEngineering (IJIRAE), 2014, Vol.1 Issue 3.

Froud Hanane, Lachkar Abdelmonaime and Ouatik Said Alaoui, “Arabic Text Summarization Based On Latent Semantic Analysis To Enhance Arabic Documents Clustering”,International Journal of Data Mining & Knowledge Management Process (IJDKP), 2013, Vol.3,No.1.

Ragunath R. And Sivaranjani N., “Ontology Based Text Document Summarization SystemUsing Concept Terms”, ARPN Journal Of Engineering And Applied Sciences, 2015, Vol. 10, No. 6.

Patil Pallavi D, Mane P M, “A Comprehensive Review on Fuzzy Logic & Latent Semantic Analysis Techniques for Improving the Performance of Text Summarization”, International Journal of Advance Research in Computer Science and Management Studies, 2014, Vol. 2, Issue11, pg. 476-485.

Dixit Rucha S., Apte S. S., “Improvement Of Text Summarization Using Fuzzy Logic BasedMethod”,IOSRJournalOfComputerEngineering(IOSRJCE)ISSN:2278-0661,

ISBN:2278-8727,2012,Vol.5,Issue 6,PP05-10.

Fachrurrozi M., Yusliani Novi, and Yoanita Rizky Utami, “Frequent Term based Text Summarization for Bahasa Indonesia”, International Conference on Innovations in Engineeringand Technology Bangkok (Thailand), 2013.

Babar S.A. and Thorat S.A., “Improving Text Summarization using Fuzzy Logic & LatentSemantic Analysis”, International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014, Vol. 1 Issue 4.

Megala S. Santhana, Kavitha A., Marimuthu A., “Enriching Text Summarization using Fuzzy logic”, International Journal of Computer Science and Information Technologies, 2014,Vol. 5 Issue 1**.**

GanesanKavita,ZhaiChengXiang andHanJiawei, “Opinosis: AGraph-BasedApproach toAbstractive Summarization of Highly Redundant Opinions”, Proceedings of the 23rd International Conference on computational Linguistics (Coling 2010), 2010, pages 340–348, Beijing.

Raymond J. Mooney and Un Yong Nahm, September 2003,“Text Mining with Information Extraction.”, Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.

“Raymond J. Mooney and Razvan Bunescu, june 2005, "Mining Knowledge from Text Using Information Extraction”, Newsletter ACM SIGKDD Explorations Newsletter - Naturallanguage processing and text mining, Volume 7 Issue 1.

“Vishal Gupta, Gurpreet Singh Lehal, AUGUST2010 “ASurvey ofText Summarization Extractive Techniques”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3.

“Sofien Lazreg, June 2012, “Using Information Extraction and TextClassification in an Effort to SupportSystematic Literature Reviews” .

“C.P. Sumathi , T. Santhanam and G. Gayathri Devi, August 2012, A SURVEY ON VARIOUS APPROACHES OF TEXT EXTRACTION IN IMAGES.” , International Journal of ComputerScience & Engineering Survey (IJCSES) (IJCSES) Vol.3, No.4,

“Anita .R. Kulkarni , Dr Mrs. S.S.Apte, Mar.-Apr. 2013, “An automatic Text Summarizationusingfeaturetermsforrelevancemeasure”,IOSRJournalofComputer

Engineering(IOSR-JCE)e-ISSN:2278-0661,p-ISSN:2278-8727Volume9,Issue3.

“Ayush Agrawal , Utsav Gupta, November 2014, “ Extraction based approach for text summarizationusingk-meansclustering”InternationalJournalofScientificandResearch Publications, Volume 4.

Story of two friendsAccessed date 16/7/2022, https://www.moralnarratives.org/the-bear-and- the-two-friends/

Story of Accessed date 19/08/2022, https://parenting.firstcry.com/articles/top- 15-shortmoral-narratives-for-kids/

N.K. Gundu1, S.M. Jadhav, T.S. Kulkarni, A.S. Kumbhar, December 2014 “ Text Extraction from Image and Displaying its Related Information” , **ISSN (Online): 2319-7064 ,** International Journal of Scientific and Research Publications, Volume 4, Issue 12, ISSN 2250-3153.

PUBLICATIONS:

1. Sawane,D.V.,Azade,S.,Khan,S.N.,Pathan,S.,Gaikwad,S.,&Gaikwad,J.(2023).

Evaluation and Techniques of Automatic Text Summarization. In Information Systems for Intelligent Systems: Proceedings of ISBM 2022 (pp. 545-557). Singapore: Springer Nature Singapore.

2. Sawane, D. V., Azade, S. Y., Khan, S. N., & Pathan, S. (2023, May). Extractive Manuscript Summarization of Motivational Blogs in Natural Language Processing. In International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022) (pp.802-809). Atlantis Press.

3. Sawane, D. V., Azade, S. Y., Gaikwad, S., Khan, S. N., Bhalerao, A., Gaikwad, J., ... & Qureshi, I. (2023, July). Retreating Nouns for Better Feature Extraction for Summarization: In Context with Narratives. In 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-6). IEEE.

4. Sawane, D. V., Azade, S., Gaikwad, S., Qureshi, I., Gaikwad, J., & Khan, S. N. (2023, November). Implementing Machine Learning Strategies to Evaluate AnalyticalAlgorithm for Unsupervised Narratives Synthesis. In 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.

5. Sawane, D. V., Azade, S., Khan, S. N., Pathan, S., Gaikwad, S., & Gaikwad, J. (2023). Novel Technique for Generating Questions from marathi language Resources. In Journal of Information and Computational Science, Volume $6^{t,}$ DOI: 10.12733.JICS.2023.VI3I6.535569.79725.(Pg. No. 254-248).

6. Pathan, S., Azade, S. Y., Sawane, D. V., & Khan, S. N. (2023, May). Hyperspectral Image Classification: A Review. In International Conferenceon Applications ofMachine Intelligence and Data Analytics (ICAMIDA2022) (pp. 582-591).Atlantis Press.